ParaSuite: Boosting LLM Reasoning via Paradox Resolution

Bin Chen*[1], Yu Zhang*[1], Hongfei Ye*[1], Huiyang Wang[1], Wenxi Liu[1], Hongyang Chen[2]†

[1]University of Chinese Academy of Sciences, [2]Zhejiang Lab

{chenbin232,zhangyu2312,yehongfei23,wanghuiyang23,liuwenxi23}@mails.ucas.ac.cn, dr.h.chen@ieee.org

Abstract

Logical reasoning is a key capability of large language models, yet current benchmarks focus almost entirely on tasks that just check basic logical consistency and overlook the reflective reasoning required for paradox detection and resolution. To fill the gap, we present ParaSuite, the first pipeline dedicated to paradox research that automates data synthesis, evaluation, and training. We introduce PARADOX, a synthetic, high-quality data spanning two difficulty tiers and three academic domains, accompanied by specialized evaluation metrics and solving algorithms. We propose ParadoxBreaker-7B, trained with Mutual-Information Guided Fine-Tuning and reinforcement learning step verify paradox reward(PAPO). Experiments demonstrate significant improvements in both paradoxical and general STEM reasoning.

1 Introduction

Logical reasoning is a core capability for large language models (LLMs), yet almost all existing benchmarks—and the training objectives they induce—remain confined to classical predicate-logic consistency. Although recent advances such as chain-of-thought (CoT) prompting(Wei et al., 2022), supervised finetuning(Zhou et al., 2025), and reinforcement learning(Xie et al., 2025) have improved deductive performance, LLMs still operate almost exclusively within the safe bounds of standard syllogistic reasoning.

Real-world inference is far richer: it is often non-monotonic, context-dependent, and propositionally tangled. Paradoxes epitomize this complexity. A paradox arises when assuming a statement to be either true or false yields, via valid reasoning, a conclusion that contradicts the very assumption. Classic examples—Russell's set paradox in mathematics, Zeno's arrow in physics, or legal dilemmas in jurisprudence—have historically driven scientific and societal progress precisely by exposing cracks in orthodox logic.

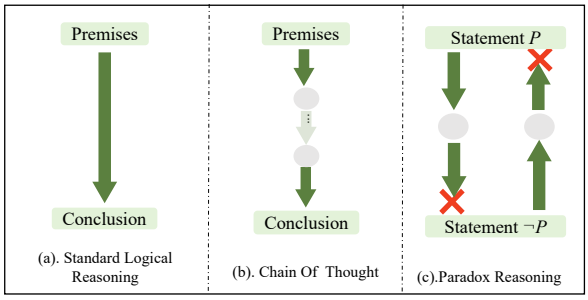

Figure 1: Three Reasoning Paradigms: Standard Logical, Chain-of-Thought, and Paradox Reasoning

Paradox proof
Input: Logical paradox problem P.
Output: Two proofs are used to discover the contradiction of the proposition P
- Suppose P is true, prove P to be false;
- Suppose P is false, prove that P is true;
- determine whether the paradox P is contradictory.

Paradox resolution
Input: Logical paradox problem P.
Output: Meta logical analysis.

Figure 2: Our setup for the paradox resolution task.

Our empirical analysis indicates that even the most widely used, reasoning-strong LLMs still struggle to recognize, explain, or resolve such contradictions. Hence, to push models beyond mere consistency checking and toward deeper reflective inference, we must move past predicate-logic benchmarks and introduce dedicated paradox-reasoning tasks and objectives(Han et al., 2024; Zhang et al., 2025).

Logical reasoning in large language models (LLMs) has received increasing attention (Zhou et al., 2024; Zhang et al., 2024; Hao et al., 2024; Xu et al., 2025), driving the development of benchmarks such as LogicBench (Parmar et al., 2024), Multi-LogiEval (Patel et al., 2024), and CriticBench (Lin et al., 2024). These benchmarks focus

*Equal contribution. The first three authors contributed equally to this work. †Corresponding author: Hongyang Chen (dr.h.chen@ieee.org).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 44246–44256

mainly on accuracy in commonsense and deductive tasks, often neglecting the reasoning process itself. Recent advances have introduced prompting methods like CoT, Self-Consistency (Wang et al., 2022), and Tree-of-Thoughts (Yao et al., 2023) to improve multi-step inference. Fine-tuning on structured datasets (e.g., ProofWriter (Tafjord et al., 2021)) and reinforcement learning approaches such as RLHF (Ouyang et al., 2022) further enhance LLMs' reasoning capabilities. Hybrid methods also support scientific and domain-specific reasoning (Lewis et al., 2020; Zhou et al., 2023; Wang and Shi, 2025; Jiang et al., 2025). However, most existing work centers on predicate logic, with little attention to paradoxical reasoning. Propositional paradoxes—though conceptually challenging—are underrepresented due to difficulties in construction and evaluation (Tian et al., 2021; Xu et al., 2025; Fan et al., 2024), and current benchmarks lack dedicated datasets or metrics to assess this critical dimension. The absence of specialized datasets, assessment tasks, and evaluation metrics has led to a lack of systematic evaluation of large language models' capabilities in this critical area.

To address this gap, we present ParaSuite, the first pipeline dedicated to paradox research that automates data synthesis, evaluation, training methods. Specifically, we introduce a comprehensive paradox dataset comprising paradoxes from mathematics, physics, and semantic logic. These paradoxes are carefully curated and categorized based on their logical structure and domain-specific characteristics. The dataset is designed not only to challenge models' ability to follow chains of reasoning, but also to push the boundaries of their capacity to detect and analyze contradictions—an essential but often neglected aspect of robust reasoning. Building upon this dataset, we fine-tuned large language models with two specific goals: (1) to enhance their ability to recognize and explicitly identify potential contradictions within paradoxical statements; and (2) to endow them with "meta-logical" explanation skills—that is, the capability to provide abstract or high-level commentary when they are unable to resolve the contradiction through standard reasoning alone. This meta-logical perspective allows models to articulate the nature of the paradox and the structural source of contradiction, reflecting a more nuanced understanding than binary logical consistency. Our experiments demonstrate not only improved paradox handling, but also reveal an unexpected auxiliary benefit: models fine-tuned on paradox resolution exhibit enhanced reasoning and explanation abilities across broader STEM (science, technology, engineering, and mathematics) domains. For example, models show improved consistency when confronted with ambiguous or contradictory premises in math word problems, physics scenarios, and logical puzzles.

In summary, our study contributes:

- We present ParaSuite, the first pipeline dedicated to paradox research that automates data synthesis, evaluation, training methods and model ParadoxBreaker-7B.
- We diagnose the state-of-the-art LLM deficiencies in paradox analysis and resolution, introduce a structured paradox dataset, and designed evaluation metrics for paradox analysis and resolution. Extensive experimental results demonstrate that our dataset is of significant value in enhancing both the logical reasoning abilities of large language models and their capabilities in the natural sciences.
- We design a step-level verifiable reinforcement-learning approach (PAPO) that trains the model to detect and resolve paradoxes effectively. Experiments show that the same method also improves large language models on other STEM reasoning tasks.

2 Related Work

Reasoning with LLMs. Large language models (LLMs) have demonstrated impressive capabilities across a broad range of complex reasoning tasks; however, challenges remain in structured forms of inference such as deductive and abductive reasoning (Clark et al., 2020; Saparov and He, 2022). Prompting strategies—including CoT and its variants like Self-Consistency (Wang et al., 2022) and Tree-of-Thoughts (Yao et al., 2023)—have consistently improved logical reasoning performance by promoting stepwise, multi-stage problem solving. Beyond prompting, supervised fine-tuning on datasets such as ProofWriter (Tafjord et al., 2021) and LogiQA (Liu et al., 2020), as well as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), further enables LLMs to internalize formal reasoning patterns and better align their outputs with human expectations. For scientific and domain-specific reasoning, LLMs must also integrate external knowledge and causal logic.

Approaches such as retrieval-augmented generation (Lewis et al., 2020) and zero-shot CoT prompting (Kojima et al., 2022) provide performance improvements without requiring model retraining. Additionally, hybrid methods that combine LLMs with external tools have proven effective in specialized tasks, including chemical synthesis planning (Zhou et al., 2023) and robotics (Huang et al., 2022). Altogether, these advancements illustrate promising progress, yet robust and interpretable logical reasoning in LLMs remains an open research frontier.

Paradox Reasoning. Tennant (2024) established a proof-theoretic framework for defining paradoxes as derivations that resist normalization, using this approach to analyze and categorize intensional paradoxes. To tackle paradoxical reasoning, Tong et al. (2023) proposed a framework that integrates inference with planning, providing a structured method to reduce inconsistencies. Additionally, Tian et al. (2021) introduced LogicNLI, a dataset aimed at evaluating first-order logical reasoning (FOL) in paradoxical contexts, while Kazemi et al. (2024) developed BoardgameQA, which included contradictory information to assess reasoning capabilities. However, existing research often lacks dedicated datasets and tailored evaluation metrics specifically for paradox reasoning, and there is a deficiency in the systematic evaluation and analysis of advanced LLMs' capabilities. This paper seeks to fill these gaps by presenting a specialized dataset and custom evaluation metrics for paradox reasoning, thereby establishing a comprehensive benchmark for state-of-the-art LLMs.

3 Task Setups and Datasets

Task Setup. The overall task pipeline is illustrated in Figure 2. We define two subtasks for evaluating large language models (LLMs):

- **Paradox Contradiction Identification (PCI):** Assessing the ability of LLMs to recognize and explicitly identify potential contradictions within paradoxical statements. The PCI task consists of two components: (i) judgment, where the model determines whether a contradiction exists, and (ii) justification, where the model provides a proof or explanation for the identified contradiction.
- **Meta-Logical Explanation Generation (MEG):** Evaluating the capability of LLMs to provide abstract or high-level commentary ("meta-logical" explanations). If the model cannot resolve the contradiction within the classical logic paradigm, it must generate a meta-logical explanation. This output should discuss, in a higher-level and abstract manner, why the paradox arises (e.g., category error, semantic ambiguity, self-reference problem), and may suggest avenues for resolution beyond classical logic.

3.1 Dataset Construction.

Harvesting and Generation. We constructed our dataset through a multi-stage harvesting, filtering, and validation pipeline tailored to support both Paradox Contradiction Identification (PCI) and Meta-Logical Explanation Generation (MEG) tasks. Starting from 60 canonical paradoxes sourced from reputable resources such as Wikipedia and spanning mathematical, physical, and semantic domains, we expanded the collection via GPT-4o, instructing the model to generate novel paradoxical statements across five targeted domains. This process yielded over 4,900 raw candidates, including 2,000 semantic, 1,200 physical, and 1,700 mathematical paradoxes.

Automatic and Human Filtering. Each candidate underwent a rigorous two-stage filtering process combining both automatic and human screening. In the first stage, five human reviewers and GPT-4 collaboratively eliminated propositions exhibiting clear implausibility or triviality. In the second stage, human annotators performed fine-grained vetting, excluding statements that were logically malformed, factually incorrect, or contained sensitive ethical or political content. Surviving candidates were classified as paradoxes if—under formal evaluation by human annotators and GPT-4—they admitted two distinct derivations (one for A, one for $\neg A$) within a logical system L given a background set of assumptions E.

Annotation for PCI and MEG Tasks. To support the PCI task, each paradox is annotated with both a judgment—i.e., whether a contradiction exists—and a justification in the form of a formal proof. For the MEG task, verified paradoxes are further processed by GPT-4o to generate high-level "meta-logical" explanations, especially in cases where contradictions cannot be resolved via standard inference. Each model output is then manually evaluated by five independent annotators for correctness and solvability judgment, with inter-rater agreement (IRA) used to ensure reliability. Outputs

with low IRA are regenerated and re-evaluated until high-quality, consistent annotations are obtained. The final Paradox dataset consists of 1,917 entries, with 1,000 used for training and 917 for testing.

Adversarial Paradox Construction. To further probe model generalization, we include *adversarial paradoxes*—newly constructed variants in which surface features are altered while the underlying logical contradiction remains unchanged. This design prevents memorization and ensures that models genuinely understand paradoxical structures rather than relying on superficial pattern matching.

3.2 Evaluation Protocol

Rubric-based metrics. Figure 3 illustrates our evaluation protocol for PCI and MEG. Rather than relying on string matching, we score both tasks with rubric-based criteria that allow semantically equivalent but differently phrased answers to receive credit.

Paradox Contradiction Identification (PCI). PCI evaluates whether a model can correctly identify the existence of a contradiction and justify it through dual-branch reasoning. We assess PCI along three dimensions: *(i) judgment correctness*, namely whether the model correctly predicts the presence or absence of a contradiction; *(ii) dual-branch validity*, namely whether the model can derive contradiction under the assumption that the proposition is true and under the assumption that it is false; and *(iii) reasoning completeness*, namely whether the proof chain is logically coherent and sufficiently complete. For each branch, responses receive full, partial, or zero credit according to a task-specific rubric. To account for generation variance, we adopt a **Pass@5** protocol and retain the highest rubric score among five sampled generations for each proof branch. The final PCI score is computed as

$$\text{PCI} = 0.5\,\text{JA} + 0.25\,P_1 + 0.25\,P_2,$$

where $\text{JA} \in \{0, 1\}$ denotes judgment accuracy, and $P_1, P_2 \in [0, 1]$ denote the rubric-based scores for the two proof branches.

Meta-Logical Explanation Generation (MEG). MEG evaluates whether a model can provide a high-level explanation when a paradox cannot be satisfactorily resolved within the classical logic paradigm. We score MEG along three dimensions: *(i) paradox source identification*, i.e., whether the explanation correctly identifies the source of the paradox such as self-reference, semantic ambiguity, or category error; *(ii) explanatory adequacy*, i.e., whether the explanation meaningfully accounts for why the contradiction arises; and *(iii) resolution framing*, i.e., whether the response suggests a plausible meta-logical perspective or resolution path beyond classical inference. Each response is assigned full, partial, or zero credit under this rubric, and the final MEG score is the mean credit over all instances that require meta-logical explanation. Importantly, semantically valid alternative explanations can receive full or partial credit even when they do not follow the exact wording of the reference annotation.

Human–LLM joint adjudication. To ensure rigorous and transparent assessment, we adopt a human–LLM joint evaluation pipeline. For each model output, GPT-4 first provides an initial assessment, including contradiction judgment, supporting evidence, and brief justification for PCI, as well as preliminary relevance and clarity assessment for MEG. Subsequently, at least two human annotators independently review the output against the annotation rubric, resolve ambiguous or partially correct cases, and override the LLM judgment when necessary, always with explicit rationale. When disagreements arise either between the LLM and human annotators or between annotators themselves, a third expert adjudicates the case. All responses are scored on a three-level scale: full credit for logically sound and sufficiently complete answers, partial credit for plausible but incomplete answers, and zero credit for incorrect or irrelevant answers. Final PCI and MEG scores are reported as the mean credit across all evaluation instances. The inter-annotator agreement is high, with IRA > 0.9.

Illustrative example. Figure 3 shows a representative example. The gold label is *contradiction*, whereas the model predicts *no contradiction*, yielding $\text{JA} = 0$. For the two proof branches, `Proof 1` correctly derives a contradiction and receives full credit ($P_1 = 1.0$), while `Proof 2` fails to establish the contradiction and receives zero credit ($P_2 = 0.0$). The resulting PCI score is

$$\text{PCI} = 0.5 \times 0 + 0.25 \times 1.0 + 0.25 \times 0.0 = 0.25.$$

For MEG, the model provides a semantically adequate high-level explanation of the paradox, and thus receives full credit, yielding $\text{MEG} = 1.0$.

Mutual-Information Guided Fine-Tuning(MIFT) Using integrated gradient

Question		Is an omnipotent being capable of creating a task that it cannot complete, while still retaining its omnipotence?
Response	Paradox Contradiction Identification (PCI)	**Proof 1 (Assuming the statement is true):** If an omnipotent being can create a task it cannot complete, it contradicts the definition of omnipotence, which entails the ability to complete all tasks. This leads to a logical contradiction. **Proof 2 (Assuming the statement is false):** If an omnipotent being cannot create a task it cannot complete, it retains the ability to complete all tasks, preserving its omnipotence without contradiction. **Conclusion:** There is no contradiction.
	Meta-Logical Explanation Generation (MEG)	**Proof:** Omnipotence does not imply the power to do the logically impossible, such as making 1+1=3 or creating a stone one cannot lift. These are contradictions, not genuine tasks. **Conclusion:** Omnipotence entails the ability to do all logically possible things; self-contradictory acts fall outside its scope.

Evaluation metrics		Judgment Accuracy	Justification Proof
Indicators	Paradox Contradiction Identification (PCI)	No contradiction. (error)	**Proof 1 (Assuming the statement is true):** If an omnipotent being can create a task it cannot complete, it contradicts the definition of omnipotence, which entails the ability to complete all tasks. This leads to a logical contradiction. (correct) **Proof 2 (Assuming the statement is false):** If an omnipotent being cannot create a task it cannot complete, it retains the ability to complete all tasks, preserving its omnipotence without contradiction. (error)
	Meta-Logical Explanation Generation (MEG)	-	The omnipotence paradox is resolved by clarifying that omnipotence does not entail doing the logically impossible, since contradictions are not genuine tasks but meaningless constructs.(correct)

Figure 3: Worked example of rubric-based evaluation for PCI and MEG. The model receives zero credit for judgment correctness, full credit for one valid proof branch, zero credit for the other branch, and full credit for a semantically adequate meta-logical explanation.

based attribution we observe that a small set of tokens such as true and false is pivotal for both detecting and resolving paradoxes and that these tokens exhibit markedly higher mutual information with the reference answer. Building on this finding we propose an MI guided fine tuning scheme that reweights the training loss according to the mutual information between each token and the reference answer.

$$\mathcal{L}_{\text{MI}} = -\sum_{i=1}^{N} \omega_i \log P_\theta\big(t_i \mid t_{<i}, \mathcal{X}\big)$$

$$\omega_i = 1 + \alpha \, \frac{I\big(t_i \,;\, \mathbf{Y}\big) - \mu}{\sigma + \varepsilon}$$

Here $I(t_i; \mathbf{Y})$ denotes the estimated mutual information between token t_i and the full answer $\mathbf{Y}$; μ and σ are the mean and standard deviation of the mutual–information scores within the current sequence, and ε is a small constant for numerical stability. α is a temperature coefficient that controls the strength of reweighting.

PAPO: Reinforcement Learning with Step Verify Paradox Reward Rewarding only the final answer often invites reward hacking. We therefore pair the outcome score with a verifiable step-wise reward and train the policy with both signals within the training loop. This method systematically evaluates each model-generated completion in response to paradox prompts and returns a scalar reward that reflects both logical correctness and response structure.

For each completion, the method performs a multi-objective assessment: Contradiction denial detection serves as an outcome reward. It first checks whether the output explicitly denies the presence of a contradiction. If so, it assigns a strong negative reward, ensuring that the model is discouraged from producing evasive or incorrect judgments that undermine the core objective of paradox identification.

Proof structure recognition assigns verifiable rewards to individual reasoning steps using the SMT solver Z3. For each generated step, GPT-4o first translates the natural-language premises and conclusion into a Z3-compatible subset of SMT-LIB. Let $P_1, \ldots, P_k$ denote the original premises together with all previously verified intermediate conclusions, and let R be the conclusion of the current step. We then check the satisfiability of

$$P_1 \wedge P_2 \wedge \cdots \wedge P_k \wedge \neg R.$$

If Z3 returns `unsat`, the step is verified as logically valid; otherwise, if it returns `sat` or times out, the step is treated as incorrect or unverified. Based on this step-level verification, the reward further encourages valid proof-by-contradiction reasoning, especially when the model correctly derives inconsistency under either the assumption that the proposition is true or that it is false. Each verified contradiction step receives an additional positive reward.

The outcome of the method is a dense reward signal per output, directly used to compute policy gradients like GRPO(Shao et al., 2024) optimization process. Through iterative training, this reward function reliably steers the model to generate outputs that not only pinpoint the contradiction at the heart of each paradox but also mirror expert-level reasoning processes, thus fostering deeper logical proficiency and meta-cognitive reasoning skills in the Qwen model.

4 Experiment

4.1 Experimental Setup

All experiments are run on 8 NVIDIA A100 GPU paired with a 64–core Intel Xeon CPU under Ubuntu 20.04. We use Python 3.12 and leave both the input and output context windows at each model's default maximum length. During the training stage, we chose Qwen as our foundation because it has strong general capabilities and adaptability to professional reasoning tasks. To effectively guide the model, each training example is formatted using a structured prompt template, which consists of (1) a description of the task and (2) paradoxical propositions. All hyperparameters, such as learning rate, batch size and sequence length, are empirically optimized based on the performance of the validation set.

Datasets In addition to our PARADOX dataset, we evaluate model performance on some established benchmarks: GSM8K (Cobbe et al., 2021), ARC Challenge (Clark et al., 2018), CEval (Huang et al., 2023). GSM8K consists of 8,500 grade-school level math word problems that test multi-step arithmetic reasoning. ARC Challenge poses difficult science questions from standardized exams designed for middle and high school students, requiring external knowledge and reasoning beyond surface-level text matching. CEval is a Chinese-language benchmark containing over 13,000 exam questions across 52 disciplines from high school, university, and professional qualification tests, and serves to assess domain-specific general knowledge and reasoning in Chinese.

The human reviewers and annotators involved in dataset filtering and evaluation were research assistants. Before annotation, all participants were provided with written instructions describing the task objectives, annotation criteria, expected workload, and the intended research use of the collected annotations. They were also informed that the review process might involve logically complex, contradictory, or potentially sensitive statements, although ethically or politically sensitive content was excluded during filtering.

All annotators provided informed consent prior to participation and agreed that their annotations could be used for dataset construction, evaluation, and research publication. Participants were compensated at an hourly rate consistent with institutional standards. The study protocol was approved by the relevant institutional ethics review process.

Models We benchmark some large language models on our PARADOX test dataset (917) instances, including both closed-source APIs and open-source instruction-tuned models. The evaluated models include Qwen-2.5–72B-Base(Yang et al., 2024), Llama-3.1–70B-Base (Dubey et al., 2024), Qwen-2.5–7B-Base(Yang et al., 2024), and Llama-3.1–8B-Base (Dubey et al., 2024). These models span a wide range of architectures and scales, from compact 7B variants to frontier models with over 70B parameters. We benchmark a set of strong closed-source and open-source LLMs on the PARADOX test set (917 instances).

Model	PCI ↑	MEG ↑
GPT-o1	**77.07**	**66.90**
GPT-4	<u>69.74</u>	49.87
Qwen3-235B	69.26	52.93
GLM-4	67.81	<u>57.38</u>
Qwen2.5-72B	66.52	50.66
DeepSeek-R1	65.88	55.05
DeepSeek-2.5	65.85	48.46
Claude-3.5	64.58	53.40
LLaMA-3.1-70B	60.46	34.06
Qwen2.5-7B	60.00	42.81
LLaMA-3.1-8B	55.03	27.07

Table 1: Performance of frontier LLMs on PARADOX. Best and second-best results are marked in bold and underline, respectively.

4.2 Frontier Models on PARADOX

We first benchmark a range of frontier closed-source and open-source LLMs on PARADOX. As shown in Table 1, even the strongest models remain far from saturated, especially on MEG, indicating that paradox reasoning is still challenging for current LLMs.

4.3 Main Result

Fine-tuning and reinforcement learning substantially improve the ability of Qwen2.5 models to identify and resolve paradoxes on both PCI and MEG. Moreover, results on GSM8K, ARC-C, and C-Eval suggest that paradox-aware training can

Model	Paradox (PCI ↑)	Paradox (MEG ↑)	GSM8K	ARC-C	C-Eval
Qwen2.5-3B-BASE	0.2521	0.2275	0.7910	0.5650	0.6632
Qwen2.5-3B-MIFT	0.6025	0.3724	0.7218	0.8100	0.6750
ParadoxBreaker 3B	0.6458	0.4984	0.7527	0.8190	0.6874
Qwen2.5-7B-BASE	0.6000	0.4281	0.8540	0.6370	**0.7758**
Qwen2.5-7B-MIFT	0.7204	0.5132	0.8802	0.8545	0.7397
ParadoxBreaker 7B	**0.7984**	**0.6864**	0.9013	**0.8617**	0.7548
Qwen2.5-72B-BASE	0.6483	0.5001	**0.9150**	0.7240	–
Llama3-8B-BASE	0.5503	0.2707	0.5530	0.5930	–
Llama3-70B-BASE	0.6046	0.3406	0.7760	0.6880	–

Table 2: Comparison of model performance on paradox analysis (PCI, MEG) and general reasoning tasks. PCI evaluates paradox contradiction identification and MEG evaluates meta-logical explanation generation. Best results in **bold**, second best in underline.

Model	Semantic				Physical				Math			
	PCI	PCI_h	MEG	MEG_h	PCI	PCI_h	MEG	MEG_h	PCI	PCI_h	MEG	MEG_h
Qwen2.5-72B-Base	0.6524	0.6637	0.5540	0.5568	0.6667	0.6814	0.4697	0.4757	0.6258	0.6593	0.4767	0.4815
LLaMA3-70B-Base	0.6185	0.6212	0.3624	0.3682	0.6477	0.6585	0.3727	0.3714	0.5475	0.5663	0.2867	0.2907
Qwen2.5-7B-Base	0.5923	0.6055	0.4599	0.4574	0.6720	0.6689	0.4212	0.4233	0.5358	0.5482	0.4033	0.4047
LLaMA3-8B-Base	0.5514	0.5742	0.2683	0.2706	0.6045	0.5980	0.2939	0.3002	0.4950	0.5026	0.2500	0.2534
ParadoxBreaker-7B	**0.8734**	**0.8612**	**0.7614**	**0.7598**	**0.7630**	**0.7568**	**0.6245**	**0.6433**	**0.7589**	**0.7574**	**0.6734**	**0.6820**

Table 3: Comparison of state-of-the-art models on the Paradox reasoning dataset for Semantic, Physical, and Math domains. PCI: Paradox Contradiction Identification (automatic), PCI_{human}: human evaluation of PCI, MEG: Meta-Logical Explanation Generation (automatic), MEG_{human}: human evaluation of MEG.

also transfer to broader scientific and mathematical reasoning tasks, although the gains are not uniform across all external benchmarks.

Table 2 presents the main results on paradox analysis (PCI and MEG) together with three general reasoning benchmarks (GSM8K, ARC-C, and C-Eval). Qwen2.5-3B-MIFT is obtained by fine-tuning Qwen2.5-3B-Base on the paradox dataset, and ParadoxBreaker 3B is further trained from Qwen2.5-3B-MIFT using PAPO. Similarly, Qwen2.5-7B-MIFT is obtained from Qwen2.5-7B-Base, and ParadoxBreaker 7B is further trained from Qwen2.5-7B-MIFT with PAPO.

On the paradox reasoning benchmarks, ParadoxBreaker 7B achieves the best overall performance, reaching 0.7984 on PCI and 0.6864 on MEG, surpassing the second-best model Qwen2.5-7B-MIFT (0.7204 PCI, 0.5132 MEG). On the general reasoning benchmarks, PARADOXBREAKER 7B achieves the best score on ARC-C (0.8617), while remaining competitive on GSM8K (0.9013) and C-Eval (0.7548), although the highest GSM8K score is achieved by Qwen2.5-72B-BASE (0.9150) and the highest C-Eval score by Qwen2.5-7B-BASE (0.7758).

Comparing Qwen2.5-3B-MIFT with Qwen2.5-3B-BASE, paradox-aware supervised training yields substantial gains on both PCI and MEG, improving from 0.2521 to 0.6025 on PCI and from 0.2275 to 0.3724 on MEG. Additional PAPO training further improves performance: PARADOXBREAKER 3B reaches 0.6458 on PCI and 0.4984 on MEG. Similar trends are observed for the 7B models, where Qwen2.5-7B-MIFT improves over the base model from 0.6000 to 0.7204 on PCI and from 0.4281 to 0.5132 on MEG, while PARADOXBREAKER 7B further advances these scores to 0.7984 and 0.6864, respectively.

The transfer pattern on external benchmarks is more nuanced. On ARC-C, both MIFT and PAPO consistently improve performance, with PARADOXBREAKER 7B achieving the best score of 0.8617. On GSM8K, paradox-aware training improves the 7B model from 0.8540 to 0.8802 and further to 0.9013, but does not surpass the 72B baseline. On C-Eval, the 3B model benefits from paradox-aware training, whereas the 7B model shows a decline relative to the base model despite retaining competitive performance overall. These results suggest that paradox-aware training transfers most clearly to paradox reasoning itself and to selected external reasoning tasks such as ARC-C, while transfer to broader benchmarks is task-dependent.

Table 3 further reports PCI and MEG performance across three domains: Semantic, Physical, and Math. For each domain, we report both automatic scores (PCI, MEG) and human-evaluated

scores (PCI_h, MEG_h). PARADOXBREAKER 7B achieves the best overall performance across all domains and metrics, consistently outperforming the baselines in both automatic and human evaluation. For example, on the Semantic domain, PARADOXBREAKER 7B reaches 0.8734 on PCI and 0.8612 on PCI_h; on the Math domain, it achieves 0.7589 on PCI and 0.6734 on MEG, again exceeding all baselines. Among the baseline models, Qwen2.5-72B-BASE and Qwen2.5-7B-BASE generally outperform LLaMA3-70B-BASE and LLaMA3-8B-BASE, but still remain below PARADOXBREAKER 7B.

Averaged across systems, semantic paradoxes are the easiest, whereas mathematical paradoxes remain the most challenging, consistently yielding the lowest PCI and MEG scores. Overall, these results show that PARADOXBREAKER 7B is markedly more effective in contradiction identification and meta-logical explanation generation, with especially strong gains on paradox reasoning and selective but not universal transfer to external reasoning benchmarks.

4.4 Ablation studies

Ablation studies reinforce the effectiveness of each proposed component. Table 4 presents an ablation study to assess the impact of paradox-specific fine-tuning and PAPO training on both 3B and 7B model variants. For each model, we report performance on paradox contradiction identification (PCI), meta-logical explanation generation (MEG), and general reasoning benchmarks ARC-C.

For both 3B and 7B models, fine-tuning on the paradox dataset (`+ MIFT`) leads to substantial gains in both PCI and MEG, demonstrating the effectiveness of supervised paradox learning. Subsequent PAPO training (`+PAPO`) brings further improvements, especially for the 7B model, where PCI increases from 0.7204 to 0.7984 and MEG rises from 0.5132 to 0.6864.

Improvements in paradox resolution also translate to general reasoning domains. On ARC-C, we observe performance gains after both MIFT and PAPO: for example, Qwen2.5 7B achieves an ARC-C score of 0.9013 after PAPO, outperforming its base and fine-tuned counterparts.

These results confirm that paradox-focused training not only enhances the model's capability to identify and resolve paradoxes, but also confers broader improvements in scientific and mathematical reasoning.

Model	**PCI**	**MEG**	**ARC-C**
Qwen2.5 3B-BASE	0.2521	0.2275	0.5650
+ MIFT	0.6025	0.3724	0.8100
+ PAPO	0.6458	0.4984	0.8190
Qwen2.5 7B-BASE	0.6000	0.4281	0.6370
+ MIFT	0.7204	0.5132	0.8545
+ PAPO	0.7984	0.6864	0.9013
LLaMA3-8B Base	0.5500	0.2710	–
+ MIFT	0.6510	0.4930	–
+ PAPO	0.7130	0.6680	–
LLaMA3-70B Base	0.6050	0.3410	–
+ MIFT	0.7330	0.6930	–
+ PAPO	0.7820	0.7620	–

Table 4: Ablation study on the effects of paradox fine-tuning and PAPO training across Qwen2.5 and LLaMA3 model families. PCI, MEG, and ARC-C denote paradox contradiction identification, meta-logical explanation generation, and ARC-Challenge scores, respectively.

5 Discussion

Compared Supervised FuneTuning AND MIFT Table 5 compares standard supervised fine tuning (FT) with mutual-information guided fine tuning (MIFT) under identical data volume and training epochs for two model sizes. We report three metrics MIFT outperforms FT across both model sizes

Table 5: Performance of MIFT versus standard FT. Higher values indicate better performance.

Model	PCI	MEG	ARC C
Qwen2.5 3B Base	0.252	0.228	0.565
+ FT	0.457	0.304	0.671
+ MIFT	**0.603**	**0.372**	**0.810**
Qwen2.5 7B Base	0.600	0.428	0.637
+ FT	0.684	0.492	0.752
+ MIFT	**0.720**	**0.513**	**0.855**

and all three metrics. For the 3B model it achieves increases of 31% on PCI, 22% on MEG, and 20% points on ARC-C. For the 7B model the improvements reach 5.2% on PCI and 4.3%on MEG, together with 13.7% rise on ARC-C. MIFT not only boosts paradox performance but also yields positive transfer to the external ARC Challenge benchmark, suggesting that the mutual-information signal encourages the model to acquire more general logical representations.

Design Principles of PAPO Token-level rewards in standard GRPO frameworks are insufficient for diagnosing specific faults in the reasoning process. By modularizing the reward signal into *Contradiction Detection* (CD) and *Proof Structure*

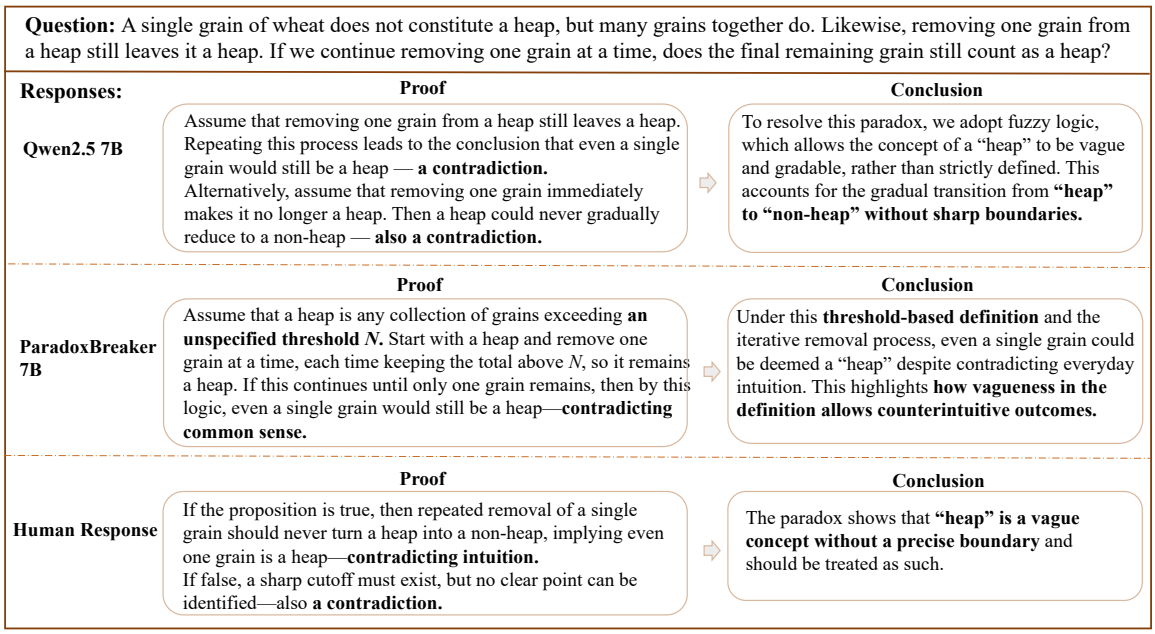

Figure 4: Cases of responses from different LLMs.

Recognition (PSR), our approach enables precise, component-level feedback that supports nuanced model guidance and diagnostics. Incorrectly denying the existence of a contradiction is more detrimental to overall task performance than providing merely incomplete proofs. Accordingly, contradiction denial errors are assigned a substantially higher negative weight in the CD component. Empirically, this asymmetric penalty structure reduced the incidence of uninformative outputs by approximately 37%.

Case Study As shown in Fig. 4, ParadoxBreaker-7B demonstrates an iterative reduction process by systematically decreasing the number of grains in the heap. This approach effectively reveals how the concept of gradation leads to ambiguity in defining a "heap." Moreover, ParadoxBreaker-7B proposes meta-logic to address the issue. In contrast, Qwen2.5-7B fails to identify the underlying contradiction of the paradox and does not provide any potential meta-logical reasoning. Overall, compared to human proofs, model-generated proofs exhibit greater vividness and logical rigor.

6 Conclusion

Current large language models reasoning largely overlook paradoxical reasoning, which often mirrors complex real-world challenges. We address this gap by constructing a comprehensive paradox dataset and explicit evaluation metrics, and propose ParadoxBreaker-7B, trained with targeted fine-tuning and reinforcement learning. Experiments demonstrate significant improvements in both paradoxical and general STEM reasoning.

Limitations

Our study is constrained by an insufficiently large dataset. Future work should aim to expand the dataset and a larger number of models to enhance the robustness of the findings.

Acknowledgment

This work is supported in part by National Key R&D Program of China (2023YFB4502400), in part by National Natural Science Foundation of China under Grant 62271452. We use an AI assistant for language translation.

References

Peter Clark, Oren Etzioni, and 1 others. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Peter Clark, Oren Etzioni, and 1 others. 2020. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, John Schulman, Jacob Hilton, and Reiichiro Nakano. 2021.

Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2024. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, and 1 others. 2024. Folio: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *Preprint*, arXiv:2404.05221.

Wenlong Huang and 1 others. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*.

Yuhui Huang, Yu Zhang, Zhengxuan Zhang, Linjun Qiu, Yuxuan Wang, Heyan Huang, and Zhengdong Hu. 2023. Ceval: A multi-level multi-subject chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. 2025. Hiddendetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2024. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems*, 36.

Takeshi Kojima and 1 others. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Patrick Lewis and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*.

Liang Liu and 1 others. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Long Ouyang and 1 others. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. *arXiv preprint arXiv:2406.17169*.

Abulhair Saparov and He He. 2022. Language models as step-by-step reasoners. In *EMNLP*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Oyvind Tafjord, Peter Clark, and 1 others. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *EMNLP*.

Neil Tennant. 2024. Which 'intensional paradoxes' are paradoxes? *Journal of Philosophical Logic*, pages 1–25.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through logicnli. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747.

Yongqi Tong, Yifan Wang, Dawei Li, Sizhe Wang, Zi Lin, Simeng Han, and Jingbo Shang. 2023. Eliminating reasoning via inferring with planning: A new framework to guide llms' non-linear thinking. *arXiv preprint arXiv:2310.12342*.

Xuezhi Wang and 1 others. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yixuan Wang and Freda Shi. 2025. Logical forms complement probability in understanding language model (and human) performance. *arXiv preprint arXiv:2502.09589*.

Jason Wei and 1 others. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning.

Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, and Wynne Hsu. 2025. Aristotle: Mastering logical reasoning with a logic-complete decompose-search-resolve framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3052–3075.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shinn Yao, Dian Zhao, and 1 others. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. 2025. Sotopia-: Dynamic strategy injection learning and social instruction following evaluation for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. *Preprint*, arXiv:2404.01230.

Jin Peng Zhou, Charles E Staats, Wenda Li, Christian Szegedy, Kilian Q Weinberger, and Yuhuai Wu. 2024. Don't trust: Verify – grounding LLM quantitative reasoning with autoformalization. In *The Twelfth International Conference on Learning Representations*.

Lujia Zhou and 1 others. 2023. Planning chemical syntheses with large language models. *Nature Machine Intelligence*.

Yujun Zhou, Jiayi Ye, Zipeng Ling, Yufei Han, Yue Huang, Haomin Zhuang, Zhenwen Liang, Kehan Guo, Taicheng Guo, and Xiangqi Wang. 2025. Dissecting logical reasoning in llms: A fine-grained evaluation and supervision study.